DOCUMENT RESUME

ED 213 753 TM 820 175

AUTHOR Yen, Wendy M.
TITLE Obtaining Some Degree of Correspondence Between Unequatable Scores: A Comparison of Item Response Theory and Equipercentile Equating Methods.
PUB DATE Mar 82
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Difficulty Level; *Equated Scores; *Latent Trait Theory; Methods; *Scoring Formulas; *Simulation; *Testing Problems; Test Reliability; Test Theory
IDENTIFIERS Data Sets; *Equipercentile Equating; *Tau Equivalence

ABSTRACT

Test scores that are not perfectly reliable cannot be strictly equated unless they are strictly parallel. This fact implies that tau equivalence can be lost if an equipercentile equating is applied to observed scores that are not strictly parallel. Thirty-six simulated data sets are produced to simulate equating tests with different difficulties and discriminations. The data are generated to produce tau-equivalent item response theory results. When an equipercentile equating is applied to these data, locally biased (that is, non-tau equivalent) results are produced for tests of unequal difficulty. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED213753

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Obtaining Some Degree of Correspondence Between Unequatable Scores: A Comparison of Item Response Theory and Equipercentile Equating Methods

Wendy M. Yen

CTB/McGraw-Hill

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. M. Yen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 175

To be presented at the American Educational Research Association meetings, New York, March 1982

Abstract

Test scores that are not perfectly reliable cannot be strictly equated unless they are strictly parallel (Lord, 1980). This fact implies that tau equivalence can be lost if an equipercentile equating is applied to observed scores that are not strictly parallel. Thirty-six simulated data sets are produced to simulate equating tests with different difficulties and discriminations. The data are generated to produce tau-equivalent item response theory results. When an equipercentile equating is applied to these data, locally biased (i.e., non-tau-equivalent) results are produced for tests of unequal difficulty.

Obtaining Some Degree of Correspondence Between Unequatable Scores: A Comparison of Item Response Theory and Equipercentile Equating Methods

Lord (1980, pp.195-198) demonstrates that fallible test scores cannot be strictly equated unless they are strictly parallel. Some of the implications of this fact are examined and related to classical true-score theory definitions of score correspondence. The practical effects of these results are explored with simulated data.

Theoretical Background

Classical True-Score Theory and Item Response Theory Definitions

Following the classical true-score theory definition (Lord & Novick, 1968, Chapter 2), let X be an observed score with $X = T + E$, where T is the true score, and E is the error score. By definition $T = \mathcal{E}(X)$, where the expectation is taken over repeated independent testings of the same examinee.

Two scores, X_1 and X_2, are tau-equivalent if

$$\mathcal{E}(X_1 | T) = \mathcal{E}(X_2 | T).$$

Two scores are parallel if

$$\mathcal{E}(X_1 | T) = \mathcal{E}(X_2 | T)$$

and

$$\sigma^2(X_1 | T) = \sigma^2(X_2 | T).$$

$\sigma^2(X|T)$ is the error score variance of X. (In popular usage, two scores are called parallel if they arise from tests that have the same average error variance taken over examinees; this more liberal definition of parallel scores will not be used here.) Two scores are strictly parallel if

$$f(X_1|T) = f(X_2|T),$$

where $f(X|T)$ is the conditional distribution of observed scores for a fixed true score (Lord, 1980, p.198). Lord & Novick (1968, p.46) call such scores equivalent.

If X_1 and X_2 are tau-equivalent, their means are expected to be the same for any group of examinees. However, their variances and higher moments taken over examinees are not expected to be equal. For example, the variance of X taken over examinees is

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 ,$$

where σ_T^2 and σ_E^2 are the true score and error score variances. If X_1 and X_2 are tau-equivalent, their true scores are equal for every examinee, and their true score variances are equal. Differences between $\sigma_{X_1}^2$ and $\sigma_{X_2}^2$ are expected depending on the values of $\sigma_{E_1}^2$ and $\sigma_{E_2}^2$. The two scores can have different correlations with other scores depending on the attenuating influences of the error scores in X_1 and X_2.

If X_1 and X_2 are parallel, their means, variances, and correlations with other variables are expected to be the same for any group of examinees, but higher order moments are not expected to be the same. Strictly parallel scores are expected to have identical distributions (and identical moments) and identical relationships with other variables.

Many of the developments in this paper arise from item response theory (IRT) as developed by Birnbaum, Lord and Novick (1968), and in particular by Lord (1980). An item response theory is defined by item response functions, $P_i(\theta)$, each of which is the probability of a correct response on item i (i = 1, ..., n) expressed as a function of item parameters for the item and examinee trait level, θ. $\hat{\theta}$ is an estimate of θ, and, when $\hat{\theta}$ is an unbiased estimate of θ, $\hat{\theta}$ is an observed score with θ as its associated true score.

Levels of Correspondence Obtained by Observed Scores

Theorem 1. Observed scores will have the same distribution for every group group of examinees if and only if the observed scores are strictly parallel.

The distribution of observed scores, X, is

$$\phi(X) = \int_{-\infty}^{+\infty} f(X|T)\ g(T)dT.$$

If the observed scores for two tests have the same distribution for a given group of examinees, then

$$\phi_1(X_1) = \phi_2(X_2),$$

and

$$\int_{-\infty}^{+\infty} f_1(X_1|T)g(T)dT = \int_{-\infty}^{+\infty} f_2(X_2|T)g(T)dT.$$

$$\frac{d}{dT}\left[\int_{-\infty}^{+\infty} f_1(X_1|T)g(T)dT\right] = \frac{d}{dT}\left[\int_{-\infty}^{+\infty} f_2(X_2|T)g(T)dT\right],$$

and

$$f_1(X_1|T) = f_2(X_2|T).$$

The converse is obviously true (Lord, 1980, p.196): if $f_1(X_1|T) = f_2(X_2|T)$, then $\phi_1(X_1) = \phi_2(X_2)$.

Theorem 1 covers the situation in which both tests are perfectly reliable, i.e.,

$$f_1(X_1|T) = f_2(X_2|T) = \begin{cases} 1 \text{ if } X = T \\ 0 \text{ otherwise} \end{cases}$$

Corollary 1.1. As the number of items becomes large, tau-equivalent maximum likelihood trait estimates will have the same distribution for every group of examinees if and only if they have the same information function.

This corollary follows from the fact that as the number of items becomes large and X is a maximum likelihood trait estimate, $\hat{\theta}$, $f(\hat{\theta}|\theta)$ approaches a normal distribution with a mean of θ and a variance of $\sigma^2(\hat{\theta}|\theta) = 1/I(\theta,\hat{\theta})$, where $I(\theta,\hat{\theta})$ is the information that $\hat{\theta}$ conveys about θ (Lord & Novick, 1968, p.457).

The following definitions are now needed: In an equipercentile equating, x_1 has an equipercentile relationship to x_2 if $F_1(x_1) = F_2(x_2)$, where F_1 and F_2 are the cumulative distributions of X_1 and X_2. If x_1 and x_2 have an equipercentile relationship and $x_1 = x_2$, the equipercentile relationship is an identity.

Corollary 1.2. The equipercentile relationship between two tau-equivalent observed scores will be an identity for every group of examinees if and only if the scores are strictly parallel.

Corollary 1.3. As the number of items becomes large, the equipercentile relationship between two tau-equivalent maximum likelihood trait estimates will be an identity for every group of examinees if and only if the two tests have the same information function.

The Importance of Tau Equivalence

Because of the labor involved in matching tests item by item, and in matching information functions, strictly parallel and parallel scores are rare. When tests are unequally difficult and a vertical equating is attempted, strictly parallel or parallel results are not possible. It is much more common for tests to be tau-equivalent. In particular it appears relatively straightforward to scale tests with IRT and obtain tau-equivalent $\hat{\theta}$ scores.

Tau-equivalence is a useful property for scores. Examinees could expect to receive the same score, on the average, from the two tests. Group means would be expected to be the same for either test, and local group means, such as school means, could be legitimately compared to the mean of a norm group. In discussing the calibration model, Angoff (1971) points out the lack of interchangeability of unequally reliable scores and makes the following statement:

> If, however, one considers the calibration model, for which the criterion of interchangeability throughout the range of scores is not intended to apply, then it may be quite reasonable to think of tests of different reliability as being "referred" to the same psychological scale, in the same sense that thermometers of different precision may be referred to the same scale of temperature. (p.568)

It appears that scores that are calibrated in Angoff's terms are tau-equivalent. Tests with unequal difficulties or standard errors cannot be expected to have equivalent scores but it is possible that they can have tau-equivalent scores.

When tests have a tau-equivalent, but no better than a tau-equivalent, relationship, examinees can expect to receive different percentiles from the distributions of the different tests. Scores on the two tests are not interchangeable, and care must be taken when they are pooled.

Equipercentile Equating and Tau-Equivalence

An equipercentile equating states that scores are "equated" or equivalent when they show an equipercentile relationship.

Corollary 1.4. If an equipercentile equating is applied to the observed score distributions of tau-equivalent scores that are not strictly parallel, the tau-equivalence can be lost.

The equipercentile equating of observed scores is not an equating of $f(X_1|T)$ and $f(X_2|T)$, and the scores equated by this method are not necessarily equivalent or tau-equivalent. For example, imagine that two tests have scores that are tau-equivalent, such as would happen if they were $\hat{\theta}$ values obtained using IRT. Imagine further that an equipercentile equating is applied to these scores and that this equating happens to be linear. The equating would find that

$$\frac{X_1 - \bar{X}_1}{\sigma_{X_1}}$$

is equivalent to

$$\frac{X_2 - \bar{X}_2}{\sigma_{X_2}}.$$

If $\sigma_{X_1} \neq \sigma_{X_2}$ (as expected when X_1 and X_2 are not parallel), any given score would appear more extreme for the test with the smaller value of σ_X, which is the more reliable test. Thus an equipercentile equating of observed scores, while having the appearance of equating the scores, can have the effect of setting equal scores that reflect different levels of performance and creating systematic biases between the scores.

Because an equipercentile equating is an identity in the percentile metric, such an equating produces strictly parallel scores in the percentile metric. However, it is rare for data analyses (e.g., calculation of means, standard deviations, distributions) to be performed on percentiles. Furthermore, an equipercentile equating does not respond to the appropriateness of the equating; any two variables can be subjected to an equipercentile equating. The benefits of having tau-equivalent percentiles are not clear.

In summary, test users cannot normally expect scores for tests of unequal difficulty or unequal reliability to be equated (equivalent) but they can be calibrated (tau-equivalent). It is reasonable to examine the quality of a calibration by examining whether calibrated scores have the same means for various groups of examinees. Successfully calibrated scores do not necessarily have the same variances or the same distributions. It is not necessarily reasonable to attempt to equate or calibrate tests of unequal difficulty or reliability by performing an equipercentile equating of observed scores, and the result of such an equipercentile equating cannot be assumed to be the criterion against which other procedures are to be judged.

Despite these theoretical conclusions, in practice there may not be important differences between results of equating methods such as the IRT and equipercentile methods. A simulation was conducted to shed some light on the practical significance of the theoretical conclusions.

Simulation

The simulation was restricted to a comparison of IRT and equipercentile results. Data were generated to be consistent with an IRT model; the effects of applying an equipercentile equating to these data were examined.

Data Generation

The three-parameter logistic model, which has the following item response function, was used:

$$P_{ik}(\theta_j) = c_{ik} + \frac{1 - c_{ik}}{1 + e^{-1.7a_{ik}(\theta_j - b_{ik})}} .$$

The parameters a_{ik}, b_{ik}, and c_{ik} are the discriminating power, difficulty, and lower asymptote of the i-th item in the k-th test and θ_j is the trait value of the j-th examinee. Tests of 20 and 40 items were simulated; the 40-item tests contained the 20 items duplicated. Base values of b_i were set at what appeared reasonably realistic values: -1.46, -1.25, -1.04, -.83, -.66, -.59, -.52, -.45, -.38, -.31, -.24, -.17, -.10, -.03, .04, .11, .32, .53, .74, .95. To simulate differences in difficulty level for two tests, constants were added to or subtracted from the b_i values for a test. To simulate equally difficult tests ($\bar{b}_2 - \bar{b}_1 = 0$), the base b_i values were used for both tests. To simulate tests of differing difficulties .5 was subtracted from the base b_i values for Test 1 and .5 was added to the base b_i values for Test 2 ($\bar{b}_2 - \bar{b}_1 = 1$). For a more extreme difference

in test difficulty 1.0 was added to and subtracted from the base b_i values ($\bar{b}_2 - \bar{b}_1 = 2$). The a_i values were set equal to 1.0. To simulate a difference in test information expected for a 50 per cent increase in test length, the a_i values for one of the two tests were set equal to $\sqrt{1.5} \doteq 1.22$. All c_i values were set equal to .20.

True trait values, θ_j, were generated from a normal distribution with a mean of 0 and standard deviation of 1 using the IMSL normal random deviate generator, GGNML (IMSL, 1979). Sample sizes of 1,000 and 2,000 were produced. It was assumed that for each simulee observed trait values, $\hat{\theta}_{jk}$, were distributed $N(\theta_j, 1/I_k(\theta_j,\hat{\theta}))$, where $I_k(\theta_j,\hat{\theta})$ was the information function predicted by the three-parameter model for θ_j and Test k with its prespecified item parameters. Two observed trait values, $\hat{\theta}_{jk}$, were generated (again using GGNML) for each true trait value to simulate scores on the two tests.

Equating Methods

The tau equivalence of IRT "equating" was compared with the tau equivalence of equipercentile equating. Each simulee, j, had $\hat{\theta}_{j1}$ and $\hat{\theta}_{j2}$, which were generated to be tau-equivalent. The equipercentile equating was superimposed on the tau-equivalent IRT equating. For the equipercentile equating, $\hat{\theta}_{j2}$ was replaced by $f_E(\hat{\theta}_{j2})$, which was the $\hat{\theta}_1$ value that shared an equipercentile relationship with $\hat{\theta}_{j2}$.

Measures of (Lack of) Tau Equivalence

To examine the tau equivalence of pairs of scores, their local equating bias was determined. The local equating bias of scores X_1 and X_2 was defined as

$$\beta(X_1, X_2|T) = \xi(X_1|T) - \xi(X_2|T).$$

To estimate $\beta(X_1, X_2|T)$, simulees were rank ordered on the basis of T (known θ_j) and grouped into five cells. The estimate of local equating bias was taken as

$$\hat{\beta}_1(X_1, X_2|T) = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{S_{X_1}} \quad \text{for cell } i,$$

where S_{X_1} was the standard deviation of X_1 taken over all simulees. Two summary measures of local equating bias were examined: (1) the maximum taken over cells of the absolute value of $\hat{\beta}_1(X_1, X_2|T)$, and (2) the mean taken over cells of the absolute value of $\hat{\beta}_1(X_1, X_2|T)$.

For the IRT "method" $X_1 = \hat{\theta}_1$ and $X_2 = \hat{\theta}_2$. For the equipercentile method $X_1 = \hat{\theta}_1$ and $X_2 = f_E(\hat{\theta}_2)$.

Results

For descriptive purposes estimated mean proportion-correct scores for the tests were obtained from

$$\bar{P}_k = \frac{1}{N}\frac{1}{n_k}\sum_{j=1}^{N}\sum_{i=1}^{n_k} P_{ik}(\hat{\theta}_{jk}),$$

where there were n_k items in test k. The ranges of values of $\bar{P}_k$ obtained for the various simulations are contained in Table 1.

Insert Table 1 about here

These ranges in $\bar{P}_1 - \bar{P}_2$ for the simulated data were compared to ranges found for real achievement tests. Two such tests were examined: Comprehensive Tests of Basic Skills, Form S (CTBS/S; CTB/McGraw-Hill, 1973) and California Achievement Tests, Form C (CAT/C; CTB/McGraw-Hill, 1977). For CTBS/S the five levels spanning grades 2 to 12 were examined; for CAT/C the nine levels spanning grades 1 to 12 were examined. The difference in mean item difficulty at the overlap grade for each pair of adjacent levels for each Reading, Language, and Mathematics subtest was calculated using standardization data published in technical reports (CTB/McGraw-Hill, 1974, 1979). The distributions of these differences appear in Table 2. Most of the real tests had differences in the

Insert Table 2 about here

.10 to .29 category, which is roughly equivalent to the $\bar{b}_2 - \bar{b}_1 = 1$ category in Table 1; there were sizable numbers of tests with differences in the other categories as well. It appeared that the simulation conditions spanned realistic ranges of test difficulties for adjacent test levels.

Examples of simulation results are contained in Figures 1 to 3, which are

Insert Figures 1, 2 & 3 about here

based on tests with 20 items, $\bar{b}_2 - \bar{b}_1 = 2$, $\bar{a}_2/\bar{a}_1 = 1$, and N = 2,000. Figure 1 contains standard error ($1/\sqrt{I(\theta,\hat{\theta})}$) curves for the two simulated tests based on their prespecified item parameters. Figure 2 displays relative frequency distributions for the $\hat{\theta}_{jk}$ values for the two tests. The differences between these two

distributions are shown from a different perspective in Figure 3, the equipercentile plot; each plotted point is a pair of theta values that has the same percentile for the two distributions.

To quantify the degree of curvilinearity in such a plot, correlation ratios (Hays, 1973, p.683) were obtained. Two correlation ratios were calculated for each plot, using each of the two sets of $\hat{\theta}_{jk}$ values as the independent variable. The "independent" $\hat{\theta}$'s were grouped in steps of .10 (e.g., $\hat{\theta}$ values from 1.01 to 1.10 fell in one group). The two correlation ratios for each plot were very similar, and their average was taken and labelled $\bar{\eta}^2$. The squared Pearson correlation was subtracted from the average correlation ratio for each plot to measure the proportion of variance shared by the two variables beyond the variance shared in a linear relationship. Values of this statistic appear in Table 3. The

---

Insert Table 3 about here

---

curvilinear relationships in the equipercentile plots tend to increase as $\bar{b}_2-\bar{b}_1$ increases and as the number of items decreases (for $\bar{b}_2-\bar{b}_1 \neq 0$). The amount of curvilinearity is about the same for $\bar{a}_1=\bar{a}_2$, $\bar{a}_2/\bar{a}_1 = 1.22$, and $\bar{a}_1/\bar{a}_2 = 1.22$. An examination of the equipercentile plots associated with the statistics in Table 3 found that the direction of the curvilinearity was always the same as that shown in Figure 3.

Tables 4 to 7 contain the summary measures of local equating bias.

Insert Tables 4 to 7 about here

Tables 4 and 5 report results for the IRT equating. They basically verify that the $\hat{\theta}$ values that were generated to be tau-equivalent were tau-equivalent for all the simulated conditions. Tables 4 and 5 provide a reference point for interpreting Tables 6 and 7, which contain the results for the equipercentile equating.

When $\bar{b}_2 - \bar{b}_1 = 0$, the equipercentile method produced less local bias than the IRT method. The values of $\hat{\theta}_{j1}$ and $\hat{\theta}_{j2}$ were generated according to a randomizing procedure that for infinitely large sample sizes should have produced perfectly tau-equivalent results. For finite sample sizes the IRT tau equivalence is expected to be less than perfect. In particular the $\hat{\theta}_1$ and $\hat{\theta}_2$ distributions could have different means, depending upon the results of the randomizing process. Given the way the equipercentile equating was performed, the $\hat{\theta}_1$ and $f_E(\hat{\theta}_2)$ distributions had to have the same mean, standard deviation, and other moments because the $f_E(\hat{\theta}_2)$ values were reordered $\hat{\theta}_1$ values. Thus, the procedures were biased in favor of finding less total bias for the equipercentile than the IRT results.

When $\bar{b}_2 - \bar{b}_1 = 0$ and $\bar{a}_2/\bar{a}_1 = 1$, there is no theoretical expectation of a loss of tau equivalence from the use of the equipercentile method, and Tables 6 and 7 reveal little local bias under these conditions. When $\bar{b}_2 - \bar{b}_1 = 0$ and $\bar{a}_2/\bar{a}_1 \neq 1$, the equipercentile method can in theory be expected to produce local bias due to differences in observed score standard deviations. When $\bar{b}_2 - \bar{b}_1 = 0$, and $\bar{a}_2/\bar{a}_1 \neq 1$, the differences in observed score standard deviations were so small as to introduce negligible local equating bias.

When $\bar{b}_2 - \bar{b}_1 \neq 0$, the equipercentile method introduced substantial local equating bias, regardless of the value of $\bar{a}_2/\bar{a}_1$. The equipercentile local equating bias increased as $\bar{b}_2 - \bar{b}_1$ increased and as the number of items decreased. The amount of local equating bias in the equipercentile equating was directly related to the degree of curvilinearity of the equipercentile plot (Table 3).

Discussion

The Theoretical Background section led to the expectation that in the simulation the equipercentile equating would not be tau-equivalent when $\bar{b}_2 - \bar{b}_1 \neq 0$ or when $a_2/a_1 \neq 1$. When the simulated tests had equal difficulties but unequal discriminations, the equipercentile method worked well; the difference in discriminations was not large enough to produce important differences in the observed score standard deviations.

When the simulated tests had unequal difficulties, the equipercentile equating produced biased results. This bias was substantial, in some cases reaching more than one third of an observed score standard deviation.

It should be noted that the curvilinear equipercentile plots have occurred with real test data (see Yen, 1982). In fact, it was the discovery of such plots that led to the research described in the present paper.

Because the $\hat{\theta}$ values were generated to be consistent with an IRT model, the IRT equating method is not actually examined. When analyzing a set of real data with IRT and equipercentile methods, one starts with response vectors, and the differences between the IRT and equipercentile methods may change with such data. The present paper indicates that simulations beginning with response vectors would be worthwhile to pursue. There are other methods that could be used to calibrate nonequivalent tests, such as true score equating (Lord, 1980) and Method-20 (Lord, 1969), and these methods also deserve comparison. Finally, the present paper deals only with the θ metric. Comparisons in the proportion-correct metric should also be made.

References

Angoff, W.M. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, Washington,D.C: American Council on Education, 1971.

California Achievement Tests, Form C. Monterey, CA: CTB/McGraw-Hill, 1977.

California Achievement Tests, Forms C and D, Technical Bulletin 1. Monterey, CA: CTB/McGraw-Hill, 1979.

Comprehensive Tests of Basic Skills, Form S. Monterey, CA: CTB/McGraw-Hill, 1973.

Comprehensive Tests of Basic Skills, Form S, Technical Bulletin No. 1. Monterey, CA: CTB/McGraw-Hill, 1974.

Hays, W.L. Statistics for the social sciences. San Francisco: Holt, Rinehart, & Winston, 1973.

IMSL Library (7th Ed.). Houston, TX: International Mathematical and Statistical Libraries, 1979.

Lord, F.M. Estimating true-score distributions in psychological testing (an impirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Laurence Erlbaum Associates, 1980.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley, 1968.

Yen, W.M. Use of three-parameter item response theory in the development of CTBS, Form U, and TCS. Paper to be presented at the National Council on Measurement in Education meetings, New York, March 1982.

Author's Address

Please address correspondence concerning this paper to Wendy M. Yen, CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA 93940.

Table 1

Ranges of Values of $\bar{P}_k$

for Simulation Data

$\bar{b}_2-\bar{b}_1$	$\bar{P}_1$	$\bar{P}_2$	$\bar{P}_1 - \bar{P}_2$
0	.65 to .66	.65 to .66	.00 to .01
1	.75 to .76	.55	.20 to .21
2	.83 to .85	.44 to .46	.37 to .40

Table 2

Percent of Adjacent
Achievement Test Levels with
a Given Range of Difference
in Mean Item Difficulty
at the Overlap Grade

$\bar{P}_1-\bar{P}_2$	Tests	
	CTBS/S	CAT/C
.00 to .09	10	40
.10 to .29	80	54
.30 & up *	10	5

* The largest difference was .37.

Table 3

Values of $\bar{\eta}^2 - r^2$ for the Equipercentile Plots

	$\bar{a}_1 = \bar{a}_2$	$\bar{a}_2/\bar{a}_1 = 1.22$	$\bar{a}_1/\bar{a}_2 = 1.22$
$\bar{b}_2 - \bar{b}_1 = 0$			
20 items			
N = 1,000	.002	.003	.002
N = 2,000	.001	.006	.002
40 items			
N = 1,000	.006	.003	.005
N = 2,000	.003	.006	.002
$\bar{b}_2 - \bar{b}_1 = 1$			
20 items			
N = 1,000	.032	.036	.039
N = 2,000	.028	.038	.025
40 items			
N = 1,000	.017	.026	.023
N = 2,000	.017	.025	.018
$\bar{b}_2 - \bar{b}_1 = 2$			
20 items			
N = 1,000	.094	.091	.113
N = 2,000	.094	.089	.093
40 items			
N = 1,000	.051	.073	.069
N = 2,000	.063	.071	.077

Table 4

IRT Method: Maximum Local Equating Bias

	$\bar{a}_1 = \bar{a}_2$	$\bar{a}_2/\bar{a}_1$=1.22	$\bar{a}_1/\bar{a}_2$=1.22
$\bar{b}_2-\bar{b}_1 = 0$			
20 items			
N = 1,000	.07	.08	.07
N = 2,000	.02	.04	.03
40 items			
N = 1,000	.05	.06	.07
N = 2,000	.03	.03	.03
$\bar{b}_2-\bar{b}_1 = 1$			
20 items			
N = 1,000	.06	.06	.05
N = 2,000	.04	.04	.02
40 items			
N = 1,000	.05	.04	.06
N = 2,000	.04	.03	.03
$\bar{b}_2-\bar{b}_1 = 2$			
20 items			
N = 1,000	.07	.07	.07
N = 2,000	.09	.05	.04
40 items			
N = 1,000	.08	.05	.08
N = 2,000	.04	.04	.04

Table 5

IRT Method: Mean Local Equating Bias

	$\bar{a}_1 = \bar{a}_2$	$\bar{a}_2/\bar{a}_1$=1.22	$\bar{a}_1/\bar{a}_2$=1.22
$\bar{b}_2-\bar{b}_1 = 0$			
20 items			
N = 1,000	.04	.04	.04
N = 2,000	.02	.03	.01
40 items			
N = 1,000	.02	.03	.03
N = 2,000	.02	.02	.02
$\bar{b}_2-\bar{b}_1 = 1$			
20 items			
N = 1,000	.04	.04	.04
N = 2,000	.02	.02	.02
40 items			
N = 1,000	.02	.03	.02
N = 2,000	.02	.02	.02
$\bar{b}_2-\bar{b}_1 = 2$			
20 items			
N = 1,000	.04	.05	.04
N = 2,000	.04	.03	.02
40 items			
N = 1,000	.03	.04	.04
N = 2,000	.03	.02	.03

Table 6

Equipercentile Method: Maximum Local Equating Bias

	$\bar{a}_1 = \bar{a}_2$	$\bar{a}_2/\bar{a}_1$=1.22	$\bar{a}_1/\bar{a}_2$=1.22
$\bar{b}_2-\bar{b}_1 = 0$			
20 items			
N = 1,000	.04	.03	.05
N = 2,000	.02	.02	.01
40 items			
N = 1,000	.03	.02	.05
N = 2,000	.02	.01	.02
$\bar{b}_2-\bar{b}_1 = 1$			
20 items			
N = 1,000	.16	.18	.18
N = 2,000	.16	.17	.13
40 items			
N = 1,000	.08	.12	.09
N = 2,000	.11	.14	.12
$\bar{b}_2-\bar{b}_1 = 2$			
20 items			
N = 1,000	.37	.44	.36
N = 2,000	.41	.37	.26
40 items			
N = 1,000	.29	.35	.30
N = 2,000	.27	.30	.27

Table 7

Equipercentile Method: Mean Local Equating Bias

	$\bar{a}_1 = \bar{a}_2$	$\bar{a}_2/\bar{a}_1$=1.22	$\bar{a}_1/\bar{a}_2$=1.22
$\bar{b}_2-\bar{b}_1 = 0$			
20 items			
N = 1,000	.01	.02	.02
N = 2,000	.01	.01	.01
40 items			
N = 1,000	.01	.01	.02
N = 2,000	.01	.01	.01
$\bar{b}_2-\bar{b}_1 = 1$			
20 items			
N = 1,000	.08	.09	.09
N = 2,000	.09	.10	.08
40 items			
N = 1,000	.05	.06	.06
N = 2,000	.06	.07	.07
$\bar{b}_2-\bar{b}_1 = 2$			
20 items			
N = 1,000	.18	.20	.19
N = 2,000	.19	.20	.17
40 items			
N = 1,000	.14	.15	.15
N = 2,000	.14	.15	.15

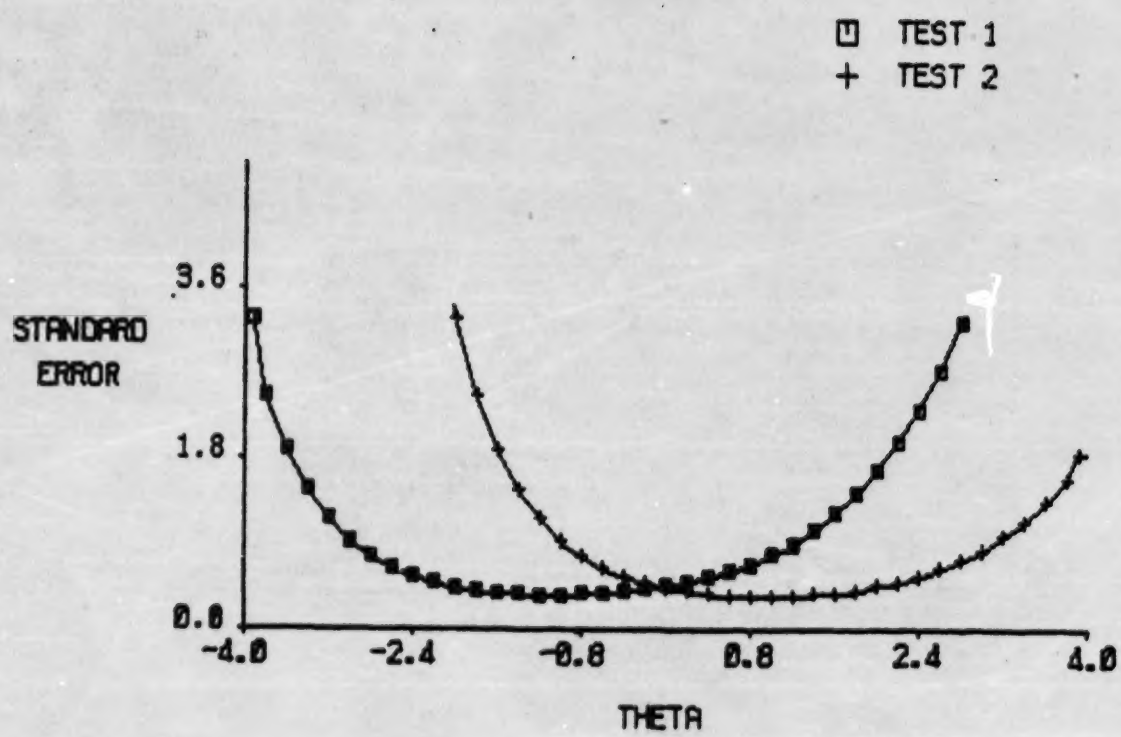

Figure 1. Standard error curves for a simulation based on 20 items, $\bar{b}_2 - \bar{b}_1 = 2$, $\bar{a}_2/\bar{a}_1 = 1$, N = 2,000.

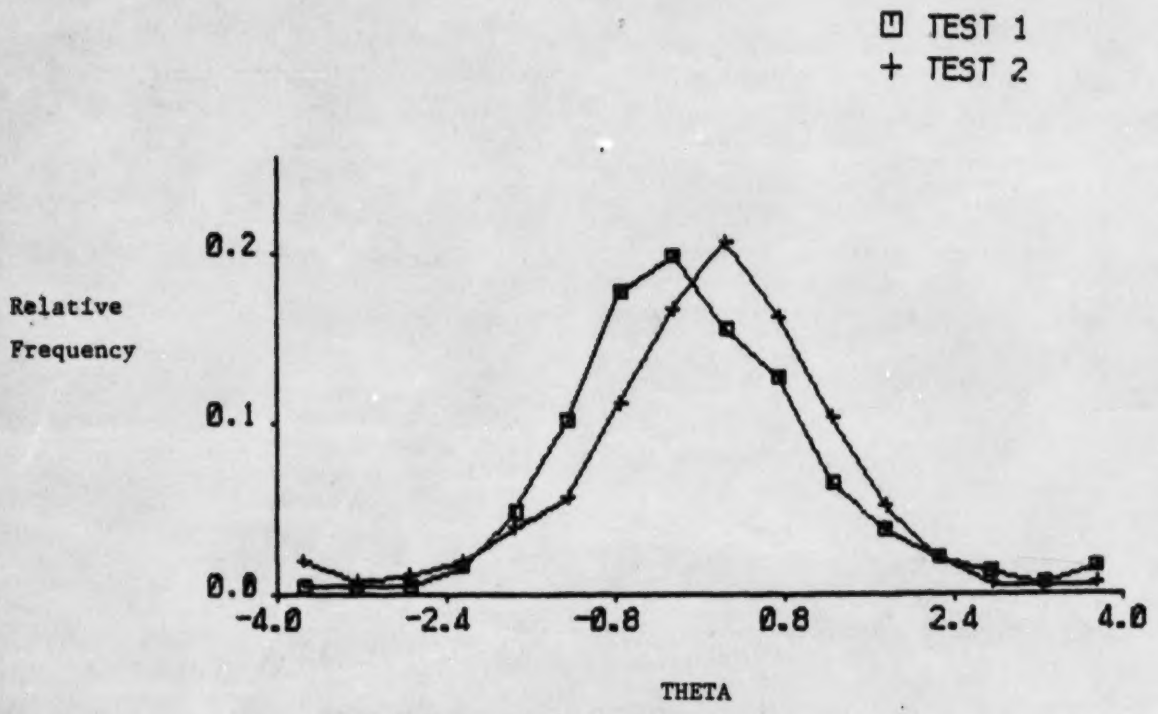

Figure 2. Relative frequency distributions of $\hat{\theta}_{jk}$ values for a simulation based on 20 items, $\bar{b}_2 - \bar{b}_1 = 2$, $\bar{a}_2/\bar{a}_1 = 1$, N = 2,000.

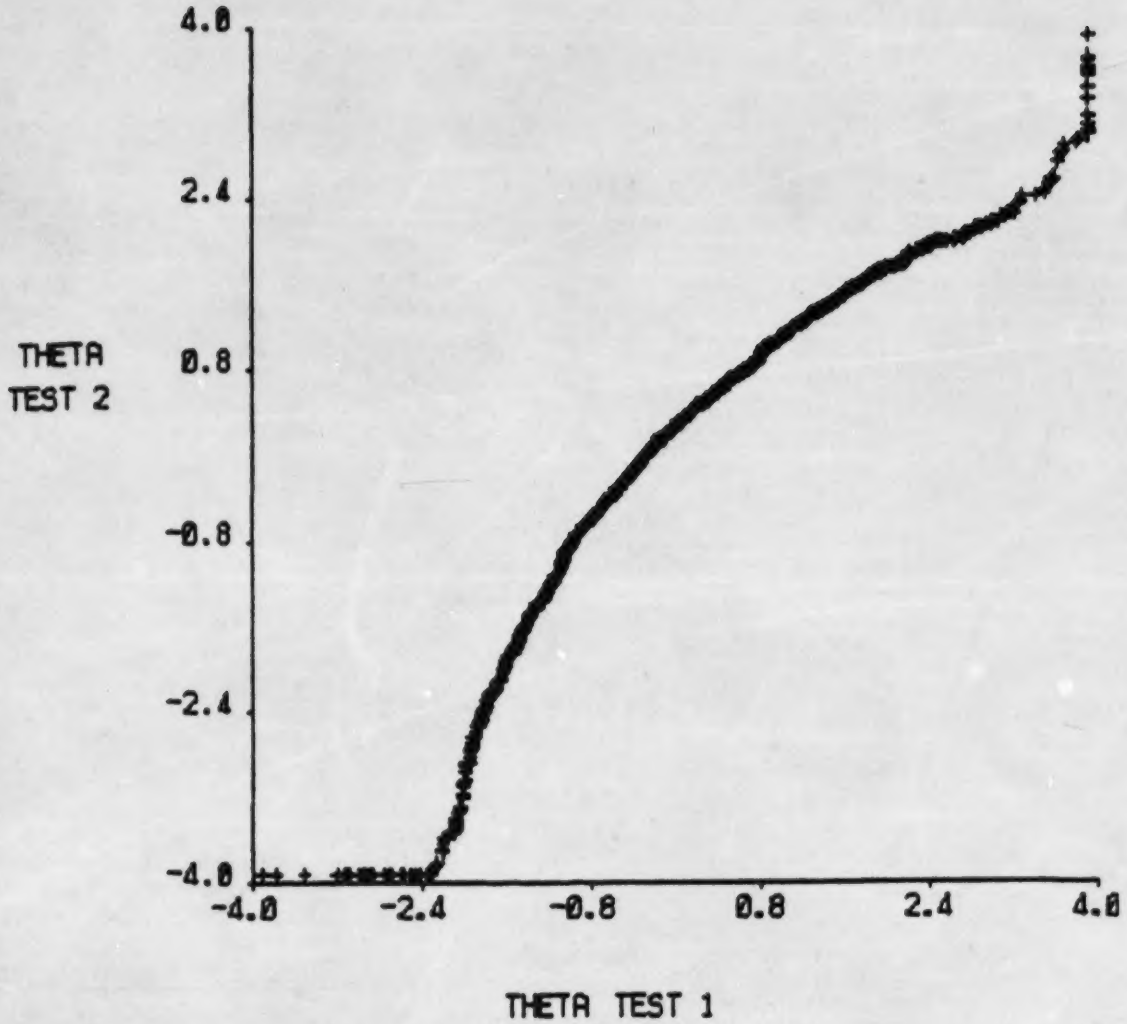

Figure 3. Equipercentile plot of $\hat{\theta}_{jk}$ values for a simulation based on 20 items, $\bar{b}_2 - \bar{b}_1 = 2$, $\bar{a}_2/\bar{a}_1 = 1$, N = 2,000.